Blockchain Mining and Analysis

Fabian Galm
University of Mannheim
1754512, MSc. Business Informatics
Germany
fgalm@mail.uni-mannheim.de

Marvin Ostermaier
University of Mannheim
1724032, MSc. Business Informatics
Germany
maosterm@mail.uni-mannheim.de

Mayank Agarwal
University of Mannheim
1725206, MSc. Business Informatics
Germany
magarwal@mail.uni-mannheim.de

José Luis Garza Garza
University of Mannheim
1725240, MSc. Data Science
Germany
jgarza@mail.uni-mannheim.de

Oghenekeno Omogha
University of Mannheim
1725331, MSc. Business Informatics
Germany
oomogha@mail.uni-mannheim.de

Masoumeh Shekari
University of Mannheim
1719408, MSc. Business Informatics
Germany
mshekari@mail.uni-mannheim.de

ABSTRACT

The public attention towards digital assets has influenced the market of Non-fungible tokens(NFTs) in achieving record sales. These NFTs are traded online mostly in cryptocurrencies, stored on the blockchain and have an associated smart contract for assigning ownership and purchase of NFTs. This paper analyzes the economic influence of the NFT space on the ethereum blockchain. We identify NFT PFP and POAP collections of interest and illustrate the data collection of NFT transaction data available on the ethereum blockchain. We study the correlation between POAP and PFP using timestamp record of transfers recorded in the blockchain, we conclude that POAP are not drivers for owning an NFT. We also structure the transaction data as a directed network graph and perform community detection to identify interactions between different NFT communities. We identify NFT POAP communities and identify that these communities interact with NFT PFP communities. Finally, we adopt a collaborative filtering approach to explore the predictiveness of economic features of NFTs as regard their social features.

CCS CONCEPTS

• **Information systems → Collaborative filtering**; **Data cleaning**; **Clustering**; **Data analytics**; **Social networks**.

KEYWORDS

Blockchain, Ethereum, Fungible Tokens, Non Fungible Tokens, Proof of attendance Protocol, Profile Picture, Non Profile Picture , Social Networks, Collaborative Filtering

1 INTRODUCTION

Blockchain technology has evolved over the years and new applications have started to emerge. One such application which uses blockchain technology is Fungible tokens (FT) and Non-fungible tokens or NFTs.

Based on ERC-20 standard, Fungible tokens are interchangeable units of commodity and are exchangeable with one another. Bitcoin is one of the most famous forms of fungible tokens in the blockchain ecosystem. On the contrary, described as a unit of digital information (token)[3], NFTs can be distinguished by their unique identifications and are traded based on different characteristics including rarity, age, liquidity, etc.[10]. It is worth to distinguish NFTs from common cryptocurrencies such as Bitcoin and Ethereum, however, NFTs are mostly registered on the Ethereum blockchain[4]. In other words, different NFT marketplaces including Opensea, leverage Ethereum cryptocurrency as a mean of payment and trading NFTs.

NFTs started increasing as of 2017 with the launch of Cryptokitties which demonstrated the economic value can be created by selling cat image tokens[8]. Since then, The NFT space has exploded which now consists of over 20,000 collections and economic value in billions. There have been projects which are used in Metaverse, PFP, POAP and digital art.

We will focus on two types of NFT projects which are present on Ethereum Blockchain: PFP and POAP. PFP (Profile picture) are NFTs which are specific to a collection and can be used as a profile picture for owners of such tokens on Social Media platforms [1] such as Twitter or Discord as a way to represent their affiliation with that NFT Collection. They often contain character like features and the most popular NFT collections include CryptoPunks, Bored Ape Yacht Club and Meebits. POAP (Proof of attendance Protocol) are NFTs which are given to attendees of an event. These events could be digital of in-presence and tokens are issued to few people attending these events by minting a NFT token particular to that event. The use-case for POAP is to record the attendance of such events in a digital format. In addition, we will also use Fungible tokens on Ethereum blockchain to observe their interaction with NFT owners and associated communities. We used Figma tool as project management, literature review and documentation tool [1]

Research Goal. Our general goal is to understand the differences between Fungible Token (FT) and NFT features by obtaining all the necessary data from the Ethereum blockchain. This will help us get insights in the NFT space, detect FT and NFT communities and finally identify correlations between FTs and NFTs based on predictiveness.

Research Questions. To achieve our goal, we want to analyze (i) the correlation between attending POAP events and owning NFTs, (ii) identify communities and their respective characteristics using transactional data and finally (iii) how predictive economic features are regarding social features.

[1]https://www.figma.com/file/hZL4mBeXmuTpRlfSVUOP7v/Blockchain-Mining-Team-project?node-id=0%3A1

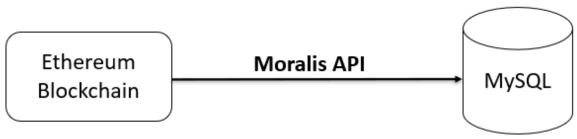

Figure 1: Data Collection Setup

2 DATA COLLECTION AND DATA SETS

The project requires various data, which can be found in the Ethereum blockchain. Therefore, an extensive collection of data takes place, in which an API is used. Another possibility would be to run a separate ethereum archival node to obtain the data. However, this requires high-performance hardware, which is not available to us. The required data from the Ethereum Blockchain is, for example, FT and NFT holders, their transactions, and their wallet balance. The API used to collect the data is the Moralis[2] API, which was chosen because of its extensive endpoints. However, the API has some limitations in terms of the requests which can be performed per second, so this restriction must be taken into account when collecting the data. Since a lot of data is needed, many API requests have to be executed, which results in significant time consumption. To avoid having to repeat the requests, the collected data is stored in a relational SQL database. This allows us to access the data quickly.

Initially, we identified over 200 NFT collection by using Opensea[3]. However, collecting data from all of these collections would be far too time-consuming, so we focused on the 10 PFP collections with the highest transaction volume.

2.1 Data Collection

In order to collect the data from the selected 10 PFP collections, different API endpoints are used depending on what data is needed. To collect the data effectively, different Python scripts were developed to automatically query the data and store it in your MySQL[4] database, as shown in figure 1.

The following endpoints of the Moralis API were used for data collection:

- GET//nft/{address}/owners
- GET//nft/{address}/transfers
- GET//{address}/erc20/transfers
- GET//{address}/erc20
- GET//{address}/nft

These endpoints provide us with with following datasets:

- NFT owner wallets
- FT and NFT transfers
- FT and NFT wallet balance

2.2 Datasets

In general, six different data sets are distinguished, whereby some of them build on each other. Each of these data sets is stored in a separate SQL table. In addition, all attributes that can be extracted from the API requests are stored.

[2]https://moralis.io
[3]https://opensea.io
[4]https://www.mysql.com/de/

2.2.1 NFT Owner Dataset. This dataset contains current NFT owners from the 10 listed PFP collections including 127,530 NFT's distributed over 47,633 unique wallets. We utilized the following information: **token_address, token_id, owner_of, name**. The unique wallets of this data set form the basis of several other datasets, which is why it is highly important. In addition, the dataset includes all POAP data, being 34,876 POAP NFTs held by 21,665 unique wallets. The unique wallets of the PFP and POAP NFTs have an overlap of 1,353 wallets.

2.2.2 FT Transfer Dataset. This dataset contains all FT transactions from 47,633 unique wallets including 7,270,851 FT transactions. We utilized the following information: **from_address, to_address, address**.

2.2.3 NFT Transfer Dataset. This dataset contains all NFT transactions from 47,633 unique wallets including 606,078 NFT transactions. We utilized the following information: **from_address, to_address, token_address, token_id**.

2.2.4 POAP Transfer Dataset. Similar to NFT transfers, this dataset contains the NFT transfers of the 1,353 wallets resulting from the overlap in the NFT Owner dataset. These are therefore the wallets that hold both a PFP from the 10 initial collections and a POAP token. This represents 3,084,853 NFT transfers from 1,353 wallets with 19,340 distinct NFT collections. We utilized the following information: **from_address, to_address, token_address, token_id**.

2.2.5 FT Balance Dataset. The FT Balance record contains the FT balances of all 47,633 wallets that are present in the NFT Owner data set. These are the FT tokens that the wallets hold at this point in time. There are 662,428 FT tokens held by the wallets, with 26,881 distinct FT tokens. We utilized the following information: **owner_of, token_address, name**.

2.2.6 NFT Balance Dataset. Similar to the FT balance dataset, the NFT balance dataset contains the NFT balances of all 47,633 wallets from the base dataset. These are 5,684,794 NFT tokens from 29,682 distinct NFT collections. We utilized the following information: **owner_of, token_address , token_id, name**.

2.3 Data Exploration: Ukraine Donation Analysis

We used the data collected above to first conduct an exploratory data analysis on Ukraine donors which helps us understand the various NFT collections which donated to Ukraine Ethereum Wallet Address. Also, we collected FT data to explore if there are users or NFT collections which donate using FT tokens and also hold PFP tokens. We also wanted to find out which organizations have donated to Ukraine donation either using the Fungible Tokens or Non-Fungible tokens.

Regarding the FT tokens donation, in table 1 the top 5 collections by ETH Volume transferred are shown with their number of transfers and how many unique wallets are donating, where Azuki was the most ETH value transferred with 12,097ETH and the most unique wallets, while Mfers has the most transfers with 2,392.

Big organizations such as Louis Vuitton or Coca Cola have raised funds to Help Ukraine using cryptocurrency donations. In table 2 we present the top 5 organizations donating to Ukraine ETH Wallet

NFT Collection	ETH Value	# Transfers	# Wallets
Azuki	12,097	22	20
Mfers	2,392	1,114	13
Meebits	164	13	13
Capsule	6.59	6	6
Clone X	4.75	9	8

Table 1: Top 5 NFT Collections donating to Ukraine ETH Wallet Addres

Address, where Peace Ukraine leads the way with 603 transfers, and then followed by Louis Vuitton with 80 transfers.

Organization	# Transfers
Peaceful Ukraine	603
Louis Vuitton	80
Coca Cola	47
Ethereum Name Service	40
Patek Philippe	29

Table 2: Top 5 Organizations donating to Ukraine ETH Wallet Address

3 RESEARCH QUESTION #1: HOW IS ATTENDING POAP EVENTS CORRELATED WITH OWNING NFTS?

The goal of this research question is to understand the POAP NFTs owners who attend the events which are held digitally and observe their awareness of NFTs. For this analysis, we use owners of PFP collections to analyse their awareness. This analysis would provide us with an insight into the which events these users seem interested in, users who are attending such events and which NFTs they already own prior to attending such events and their activity after attending.

3.1 Data Set and Data Preprocessing

To perform this analysis, we first collect the top 10 PFP collections by trading volume. We used the getnftowners and getnfttransfers endpoints to collect this data and store this data in two separate tables in MySQL Database, so to be able to query this data and find overlap of wallets. In the PFP dataset, there are overall 127,531 NFT tokens which were owned by 47,633 users.

Following are the top 10 collections ranked by their trading volume:

- `CryptoPunks`
- `Bored Ape Yacht Club`
- `Mutant Ape Yacht Club`
- `Azuki`
- `Clone X`
- `Moonbirds`
- `Doodles`
- `Meebits`
- `Cool Cats`
- `Pudgy Penguins`

Similarly, we collected the data for all the POAP NFTs which was possible due to the limited amount of NFTs. All POAP NFTs use the same contract adddress, and are distinguished by their metadata. We were able to collect all the 34,876 POAP NFTs owned by 21,665 wallets. We stored this data in different SQL tables for owners and transactions to perform the analysis. In addition, the POAP transfer data set was used to generate a transaction history for all 1,353 included wallets. Which involves in 3,084,853 NFT transfers from 1,353 wallets with 19,340 distinct NFT collections.

Our approach was to first obtain match the NFT transfers from PFP owners and POAP owners to find the overlap of wallets. We found that there were 1,353 unique wallets that occur in both datasets.

In the POAP transfers dataset, we observed that there were transfers which were not completed. This can be identified when the transfers which are made to an mint address which is identifiable by this address (0x0000000000000000000000000000000000000000). We removed these transfers as they were not completely executed. In the next step, we also performed extraction of POAP name from the metadata column in the POAP Owners to get the name of the POAP event. This was particularly challenging as the there are many attributes which are nested inside the column in the form of key-value pairs. We used a dictionary object and regex in python to store and parse this information.

3.2 Analysis

In the Analysis phase, we wanted to build a transfer history of all the transactions which have occurred based on when the POAP event occured and observe the users which have attended the events. This was possible by using the timestamp and the block number. We performed this analysis for all the 1,353 wallets. From this analysis, we were able to understand the top 10 POAP events attended by the PFP owners which are listed below in the table 3.

POAP Event	# Attendants
Bankless Member - 2021	1,213
VBW 2020 - General	1,192
VBW 2020 - Cointelegraph Official NFT	1,183
Dappcon 2019	1,007
Bankless Member - 2020	657
ETHBerlinZwei	616
Devcon5	590
The Mutant Ape's Demise	590
Beanie x Farokh Ceasefire	532
Adidas Originals: Our Future Started here	511

Table 3: Top 10 POAP Events by Number of Attendants

Similarly, we were able to get the top 10 PFP collections owned by the POAP attendees in the Table 4

3.3 Evaluation

To evaluate and conclude our analysis, We observed that we were able to find correlations between the POAP attendees and PFP owners. Furthermore, We were also able to get more insights into the NFT space. We found that users who attend POAP events, about

NFT Collection	# Tokens Hold
POAP	1,229
OpenSea Shared Storefront	1,069
Ethereum Name Service	980
Otherdeed	542
Art Blocks	540
Rarible	486
Gutter Punks - Otherside	447
adidas Originals: Into the Metaverse	347
MutantApeYachtClub	344
Gutter Punks Flyer - Meebits	294

Table 4: Top 10 NFTs Projects Attendants of POAPs Hold

92% of them already transferred an NFT before attending. This was possible using the transfer history and comparing the block number of the transfers.

We also found the average number of NFTs owned by POAP attendant to be quite high which correlates with the number of users who performed NFT transfers before attending POAP event and we found that the average number of attendees in POAP events who hold a NFT is quite high. This insight tells us that most PFP owners do not attend POAP events to get their awareness in NFT space. They mostly get informed by exchange places and other online resources or online communities which exist to build awareness in this space. The results and conclusions can be observed in the Table 5 and Figure 2.

Insight	Value
Avg. Number of NFTs owned by POAP Attendant	815
Avg. Number of POAP Events Attended	1.6
Avg. Number of NFT transfers	1397
% Wallets who performed NFT Transfers before first POAP event	92%
% Wallets who performed NFT Transfers after first POAP event	100%

Table 5: Top 10 NFTs Projects Attendants of POAPs Hold

4 RESEARCH QUESTION #2: WHICH COMMUNITIES CAN BE OBSERVED BASED ON TRANSACTIONAL DATA AND WHAT ARE THEIR PROPERTIES?

To answer this research question we examine transactional data from the listed NFT & FT Projects in order to get useful insight regarding the global network structure and its properties. By computing different community detection algorithms, different communities can be observed and will be analyzed in the subsequent steps. By investigating the generated communities, we are able to further describe the composition of the network regarding respective communities to get even more useful insights. Firstly, we preliminary explore the dataset to identify the relationship of unique NFT wallet addresses with different NFT collections using the jarccard correlation coefficient measure. Inspecting the correlation matrix

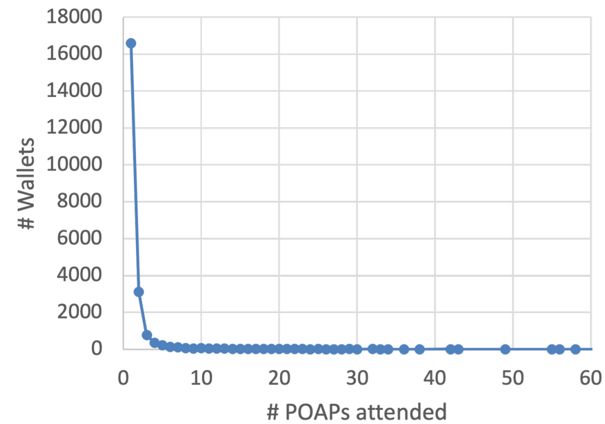

Figure 2: Attendance frequency of Wallets to POAP Events

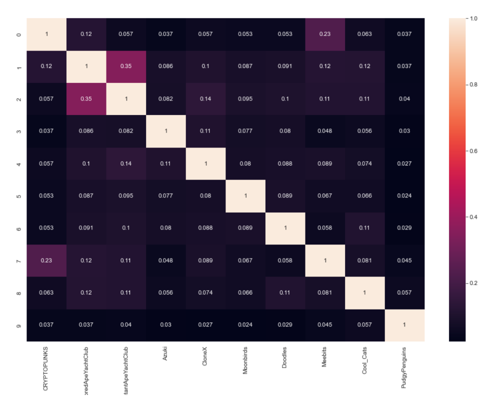

Figure 3: Correlation Matrix

in Figure 3 we infer unique walltes which are common in different collections. The figure shows that the highest correlation exists between BoreyApeYachtClub and MutantApeYacht club with a similarity score of 0,35 followed by Cryptopunks and Meebits with 0,23. For better insight, we will consider transactional data for the subsequent sections.

4.1 Data Set

As mentioned in section 2, 10 NFT projects are considered which results in 127,530 unique assets. From those assets 47,633 unique wallets are derived and are considered in the further research. As we are interested in finding communities based on transactional data, we only consider transactions between those unique wallets shown in table 6.

4.2 Analysis

To detect and analyse communities we first have to build the networks from the datasets. We use graphs for community detection since the topology of these graphs can give us valuable information about modules and hierarchical organization of these communities [6]. Hence, directed graphs have been widely used in analysing the data of social networks, where nodes represent the users and the vertices represent iterations among them[2], [5]. As our transitional data is provided with a $from\_address$ and a $to\_address$ we define our network G as a directed graph given by: $G = (V, E)$, where the vertices V are wallet addresses and the edges E are transactions. As we only consider the interactions between current NFT holders of our 10 Projects, we filter out all transactions which do not correspond to those wallet addresses.

Collection	Unique Wallets (V)	Transfers (E)
NFT	27,671	90,000
FT	13,797	78,889

Table 6: Graph properties

On filtering on the condition of current owners of NFT and FT, we obtain relevant data for building our network graphs. We observe graph properties of NFT and FT in Table 6. Furthermore, for weighting the edges in our network, we use the transaction frequency of transfers in the network. We observe the properties of the global graph for initial network analysis. We employ metrics such as *average degree* and *network density* for identifying transaction activity of wallets and the interactions of these wallet addresses respectively. Formally, the network density is defined as the "proportion of possible relationships present in a network". This statistic value ranges from **0** to **1**, with the lower limit indicating no relationships and the upper limit indicating all possible relationships in the network. Therefore, a density value closer to **0** indicates a sparse network, while a value closer to **1** indicates a dense network. On the basis of these metrics, the global network graph illustrates a sparse network with a network density of 0.00026 and average degree of 3.257. Figure 4 shows the degree distribution. It shows that there is a large number of wallet addresses with low transaction volumes and a small number with comparatively high number of transactions.

To identify the communities in our network, we adopt the Girvan-Newman algorithm. This algorithm detects communities by iteratively eliminating edges with highest shortest paths between nodes, otherwise called the edge betweeness centrality. The idea is to remove edges from the graph and break the network into smaller communities[9]. Applying this algorithm on the NFT network graph results in detection of 1190 communities, an average community size of 23.5 wallet addresses, with the highest community size of 3150 obtained from the degree of interactions with other communities in the network. We measure the quality of our detected communities using the modularity score. We obtain a modularity score of 0.716 - a higher modularity score indicates dense connection between nodes in a community.

To make assumptions about the quality of the communities, the result is compared with a null model. For this purpose, a comparatively homogeneous graph with the same properties is created and the same community detection algorithm is applied. Afterwards, the modularity score of the original graph is compared and evaluated with the randomly generated graphs. For this purpose we

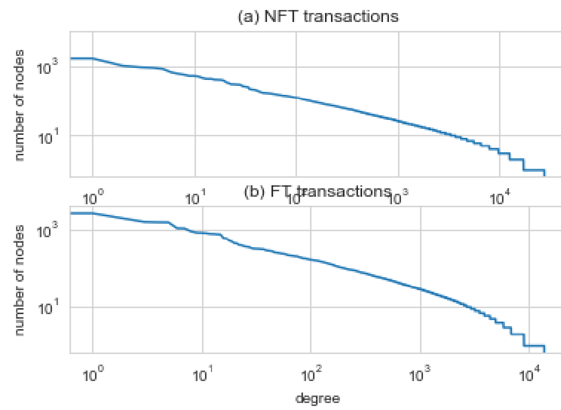

Figure 4: (a) Degree distribution of NFT transaction graph over 27,671 wallet addresses from 90,000 transactions. (b) Degree distribution of NFT transaction graph over 13,797 wallet addresses from 78,889 transactions.

created 10 random graphs and calculated the average modularity score respectively. After running 10 null models, we get an average modularity score of 0,341, which indicates a lower community quality compared to our NFT network graph.

Figure 5 visualizes the graph structure based on its communities. We can see, that many of the detected communities are separated from the giant component, so they act as their own entity. If we only consider the giant component we are left with 79 communities(6,4%) which represent 96,8% of the total number of transactions,we assume that those communities hold the majority of all the transactions. Recalculating the graph density on our giant component also reflects those findings as its now increased to 0,501.

We further investigate the interactions between the detected communities on the basis of the different collections these communities own. We identify the top 5 communities that influence and hold collections across the transaction network in Table 7.

Having detected NFT communities and their properties, we also perform community detection on the FT transactions. We also build a directed network graph $G = (V, E)$, where the vertices V are wallet addresses and the edges E are transactions using only current owners of NFT's. We also observe the properties of global network specifically the graph density, for sparsity or density of the network. The global graph density of 0.00008 indicates a sparse network. Using the Girvan-Newman algorithm we detect 516 communities with average size of 26.5 wallet addresses and a maximum community size of 1522. The quality of the detected communities is 0.709 using the modularity score measure. To make assumptions about the quality we also generate 10 null models which results in an average modularity score of 0,392. Due to the sparsity of the network, we investigate the network for the giant component in the network which contains the significant proportion of nodes in the network. We identify 56 communities that make up the giant component with a density of 0.425 and average degree of 1323.74 -

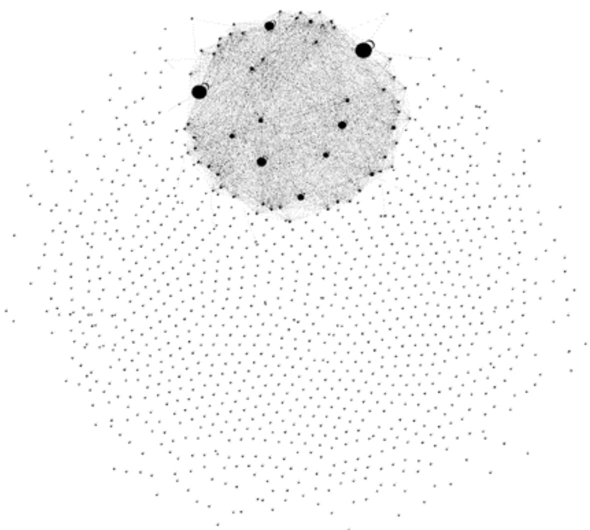

Figure 5: Graph of the community structure after clustering by communities detected by Girvan-Newman algorithm. Nodes represent communities and edges transactions.

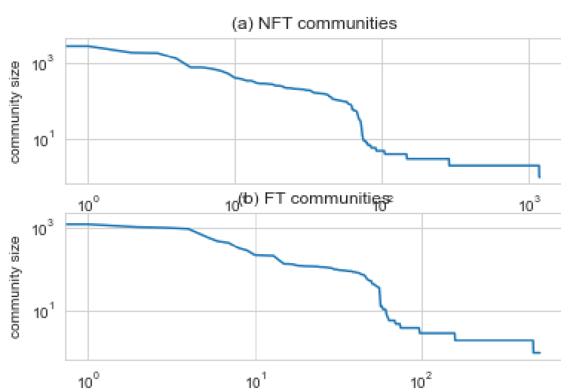

Figure 6: (a) Community size distribution on detected NFT transaction communities. (b) Community size distribution on detected FT transaction communities.

these communities are responsible for 98.33% of transactions in the network.

Given those detected communties, we can now assign each wallet address to a specific community. By inspecting the current NFT holdings for each wallet address of each community, we calculate the sum for each community respectively. Table 7 shows the top 5 communities detected on NFT transactions sorted on NFT's holdings, table 8 on FT transactions. Inspecting the tables we can see the compositions for each community and also the calculated percentage holding of all available PFP's from our 10 collections.

4.3 Evaluation

After analyzing basic properties of both FT and NFT networks, we can see that we have a high amount of wallets with a small transaction frequency and a small amount of wallets with high transaction frequency. We can assume that wallets with high frequency are more important to the network and vise versa. As we see a lot of wallets with small transaction frequency, we derive a low density for both global graphs. We detected communities and measured the quality of these communities on the metric of their modularity scores. Compared to the null model, both graphs show a higher modularity score which indicates a higher quality of the detected communities. Significantly, we identified the giant component on the NFT and FT global graph, this dense portion make up for 96.8% and 98.33% of NFT and FT transactions respectively.

5 RESEARCH QUESTION #3: HOW PREDICTIVE ARE ECONOMIC FEATURES REGARDING SOCIAL FEATURES?

The objective of this research question is to investigate the predictability of economic features in relation to social features. Here, the economic features refer to FT while the social features refer to NFT tokens. Consequently, this question examines whether implications can be derived from the FT balance data to the NFT balance data. With this in mind, a recommendation system is used that utilizes the FT balance data to recommend potential NFT projects. These recommendations can subsequently be compared to the actual NFT projects held by the corresponding wallets, which results in information about the influence of the FT balance data on the NFT projects. First, the FT and NFT balance data from all wallets in the base data set are collected. In the second step, this data is utilized in a recommendation system using user-based collaborative filtering. Finally, hyperparameter optimization is performed to obtain the best combination of hyperparameters leading to higher model accuracy.

5.1 Data Set and Data Preprocessing

Two datasets are used to investigate this research questions. These are the FT and the NFT balance dataset, as described in chapter 2. These datasets are based on the 47,633 unique wallets that appear in the NFT base dataset. However, not only 10 NFT projects are considered, but all projects owned by the wallets. Consequently, a large number of different FT and NFT tokens are included. Especially the information about which wallet holds how many FT and NFT tokens of a specific project is relevant for the research question. In order to be able to use the data for a recommendation system, different preprocessing steps are performed. First, FTs and NFTs that do not occur at least 5 times were removed. Second, wallets that do not contain at least 2 FTs and NFTs were ignored. And third, wallets that are not included in both FT and NFT balance data were removed. The result is approximately 600,000 FTs owned by approximately 40,000 wallets and approximately 5,000,000 NFTs owned by the same approximately 40,000 wallets.

Mod. Class	Azuki	BAYC	Cryptopunks	CloneX	Cool Cats	Doodles	Meebits	Moonbirds	MAYC	PudgyPenguins	PFP amount	%PFP hold
25	184	621	4797	910	406	318	6833	386	786	249	15490	12,15%
901	498	693	206	661	644	4812	503	543	834	144	9538	7,48%
415	433	315	280	424	493	161	344	333	645	3247	6675	5,23%
679	801	457	79	1057	224	297	239	456	881	131	4622	3,62%
31	208	293	131	408	1345	166	297	204	501	294	3847	3,07%

Table 7: PFP holdings of detected individual NFT transaction communities separated by collection, sorted by number of current PFP holdings.

Mod. Class	Azuki	BAYC	Cryptopunks	CloneX	Cool Cats	Doodles	Meebits	Moonbirds	MAYC	PudgyPenguins	PFP amount	%PFP hold
0	200	330	3810	736	316	231	4389	296	363	254	10925	8,56%
19	326	521	652	609	314	481	2262	1423	726	472	7786	6,10%
23	1598	392	85	1698	227	197	399	177	535	118	5426	4,25%
2	71	173	1560	309	236	104	1992	173	337	113	5068	3,97%
5	290	479	78	343	662	288	305	284	701	231	3661	2,87%

Table 8: PFP holdings of detected individual FT transaction communities separated by collection, sorted by number of current PFP holdings.

5.2 Recommender System

The recommendation system is based on a user-based collaborative filtering system to provide recommendations of NFTs based on FTs. This algorithm is popular for implementing a personalized recommendation system in the commercial environment [11]. The basic application of the algorithm is to predict user preferences based on the historical data collected from the behavior of previous "similar" users[7].

For this purpose, the FT balance data is used in the first step to generate a Wallet FT matrix, which indicates in its cells how many FT tokens of a particular project a particular wallet holds as shown in figure 7. For example, $Wallet_1$ owns three FT_1 tokens.

	FT_1	FT_2	FT_3	FT_4	FT_5
Wallet_1	3	4	2	0	0
Wallet_2	0	1	0	1	0
Wallet_3	2	0	1	2	3
Wallet_4	2	3	0	5	4
Wallet_5	0	2	3	0	1

Figure 7: Wallet-FT Matrix based on the FT Balance data set

In the second step, the similarity of the wallets is determined by using cosine similarity. This results in a matrix that indicates how similar the different wallets are as shown in figure 8. This similarity can vary from 0 to 1, where 1 corresponds to a perfect similarity. For example, the figure shows that $Wallet_1$ and $Wallet_2$ have a similarity of 0.8, which indicates that these wallets hold a similar number of similar FT tokens. The similarities along the diagonal of the wallet similarity matrix are always 1, since the same wallets are considered here. Therefore, these values must be neglected when seeking similar users in the context of user-based collaborative filtering.

In the third step of the user based collaborative filtering approach, similar users are retrieved, whereby the similarity is in terms of FTs is utilized. Consequently, the wallet similarity matrix is used. The decisive factor here is that only the similarity of all wallets to

	Wallet_1	Wallet_2	Wallet_3	Wallet_4	Wallet_5
Wallet_1	1	0.7	0.3	0.1	0.8
Wallet_2	0.7	1	0.2	0.4	0.6
Wallet_3	0.3	0.2	1	0.1	0.5
Wallet_4	0.1	0.4	0.1	1	0.8
Wallet_5	0.8	0.6	0.5	0.8	1

Figure 8: Wallet similarity matrix based on the Wallet-FT Matrix

a specific wallet is considered in order to make recommendations for this specific wallet. For this purpose, the similarity to $Wallet_1$ is considered in the following examples. To determine the similar wallets there are two different approaches. The first approach is called the "Top k" method, where a certain number k of similar wallets are selected. The k wallets with the highest similarity from the wallet similarity matrix are used. Figure 9 describes this method, using the "Top k" method with $k = 2$. The yellow columns indicate the "Top 2" wallets in terms of similarity to $Wallet_1$.

	Wallet_1	Wallet_2	Wallet_3	Wallet_4	Wallet_5
Wallet_1	1	0.7	0.3	0.1	0.8
Wallet_2	**0.7**	1	0.2	0.4	0.6
Wallet_3	0.3	0.2	1	0.1	0.5
Wallet_4	0.1	0.4	0.1	1	0.8
Wallet_5	**0.8**	0.6	0.5	0.8	1

Figure 9: "Top k" method for selecting similar wallets with $k = 2$

The second approach involves the use of a similarity threshold t, which is why this method is called the "Similarity Threshold" method. Here, the similar wallets are selected by considering their similarity value. If the similarity value of a wallet is greater than or equal to the threshold, this wallet is relevant for the recommendation. Consequently, there is no fixed number of wallets that is used, but the number can vary depending on the wallet and the threshold value t. In Figure 10, this method is illustrated using a threshold value of $t = 0.3$, which is why $Wallet_2$, $Wallet_3$ and $Wallet_5$ are chosen in terms of similarity to $Wallet_1$.

	Wallet_1	Wallet_2	Wallet_3	Wallet_4	Wallet_5
Wallet_1	1	0.7	0.3	0.1	0.8
Wallet_2	0.7	1	0.2	0.4	0.6
Wallet_3	0.3	0.2	1	0.1	0.5
Wallet_4	0.1	0.4	0.1	1	0.8
Wallet_5	0.8	0.6	0.5	0.8	1

Figure 10: "Similarity Threshold" method for selecting similar wallets with $t = 0.3$

The fourth step of the user based collaborative filtering approach to determine NFT recommendations based on FT balance data refers to the calculation of relevance values of NFTs. For this purpose, a Wallet-NFT matrix is used, which indicates how many NFTs of a specific project a specific wallet owns, as shown in Figure 11. For example, $Wallet_2$ holds three NFT_1 tokens. Formula 1 is used to calculate the relevance scores.

	NFT_1	NFT_2	NFT_3	NFT_4	NFT_5
Wallet_1	2	1	4	0	1
Wallet_2	3	4	5	0	2
Wallet_3	1	5	0	3	2
Wallet_4	2	0	1	2	4
Wallet_5	4	3	1	0	1

Figure 11: Wallet-NFT Matrix based on the NFT Balance data set

$$r\text{-}score(w_r, n_r) = \sum_{w} sim_{w_r}(w) \cdot count_{n_r}(w) \tag{1}$$

Equation 1 describes the calculation of the relevance score of a specific NFT n_r for a specific wallet w_r where w is all wallets that have been selected as similar to wallet w_r. $sim_{w_r}(w)$ denotes the similarity value of wallet w_r and wallet w, while $count_{n_r}(w)$ denotes the number of NFT tokens of a specific NFT project n_r held by wallet w. Therefore, the relevance score of a specific NFT project for a specific wallet is calculated by weighing the number of NFT project tokens owned by all similar wallets by their wallet similarity. An example of this weighting is represented by equation 2, where the "Top k" with $k = 2$ is used for the selection of similar wallets.

$$r\text{-}score(Wallet_1, NFT_1) = (0.7 \cdot 3) + (0.8 \cdot 4) = 5.3 \tag{2}$$

In the fifth and last step, the relevance scores are determined for all NFTs that the selected wallets from the similarity matrix own. This results in a list of relevance scores as shown in Figure 12, which is used to give a certain number of recommendations r.

	Relevance
NFT_1	5.3
NFT_2	5.2
NFT_3	4.3
NFT_4	0
NFT_5	2.2

Figure 12: NFT Relevance Scores for $Wallet_1$ using the relevance equation

Whereby the NFTs with the highest relevance scores are returned in descending order.

The recommendation system is thus able to provide NFT recommendations based on the FT balance data of the wallets and thus implements a user-based collaborative filtering approach.

5.3 Evaluation

To evaluate the recommendation system, first a hyperparameter optimization is performed to determine the parameters that achieve the best accuracy. For this purpose, the two methods for selecting similar wallets with different parameters are compared. For the "Top k" methods the parameters $k \in [10, 50, 100, 200]$ were used while for the "Similarity Threshold" method the parameters $t \in [0.1, 0.2, 0.3]$ were used. Furthermore, both methods were evaluated with all parameters for varying numbers of recommendations $r \in [5, 10, 20, 100]$. For the evaluation of the hyperparameter optimization the hit ratio is used as a measure of accuracy. It describes the proportion of hits to the maximum possible hits, where the number of hits is referred to as the hit rate. The hit rate indicates the number of matches between the recommendations of the recommendation system and the actual NFT projects that a wallet owns.

Figure 13 shows the results of the hyperparameter optimization for the "Top k" method, where it is evident that the hit ratio increases as k increases. Consequently, the use of more similar wallets provides a higher accuracy. This is confirmed across all numbers of recommendations, with a lower number of recommendations generally leading to a higher hit ratio.

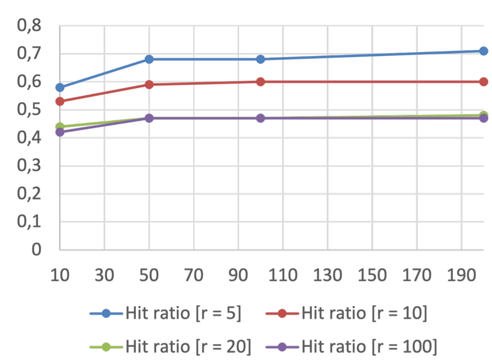

Figure 13: Hit ratio for different number of k using the "Top k" method

	5	10	20	100
Random recommender	0.01	0.04	0.07	0.33
UB-CF recommender	3.57	5.93	8.36	13.61

Table 9: Hit rate for different numbers of recommendations r using the best hyperparameters

	5	10	20	100
Random recommender	0	0	0	0.01
UB-CF recommender	0.71	0.6	0.48	0.47

Table 10: Hit ratio for different numbers of recommendations r using the best hyperparameters

Figure 14, on the other hand, shows the results of hyperparameter optimization for the "Similarity Threshold" method. It is noticeable that the accuracy increases with a higher threshold value t. Consequently, the inclusion of more similar wallets also leads to higher accuracy. However, this method generally performs better with a higher number of recommendations. Overall, it is noticeable that the "Top k" method provides better results in terms of hit ratio than the "Similarity Threshold" method. The best hyperparameters are therefore $k = 200$ using the "Top k" method for all numbers of recommendations.

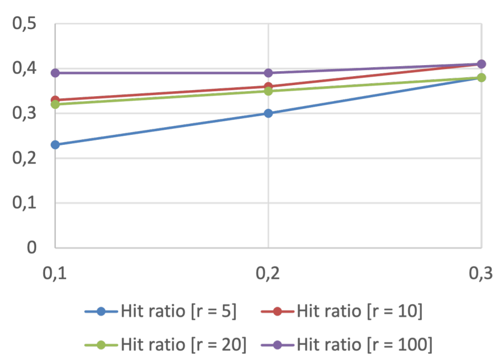

Figure 14: Hit ratio for different number of t using the "Similarity threshold" method

To evaluate the general accuracy of the recommendation system with its best hyperparameters, it is compared with random recommendations. Both the average hit rate and the average hit ratio are compared. For this purpose, 10% of the wallets of the base data set are used as a test set.

Tables 9 and 10 show the hit rate and hit ratio, respectively, for the two recommendation systems across various numbers of recommendations. It is noticeable that the user-based collaborative filtering recommendation system provides significantly higher accuracy than random recommendations. Especially for a low number of recommendations the user based collaborative filtering system provides a high hit ratio. In general, it can be said that the selected approach provides good results and an increasing number of similar wallets considered favors increases the accuracy. This can be explained by the fact that more similar wallets are taken into account. The recommendation system thus shows that there is a correlation between the economic features in the form of FTs and the social features in the form of NFTs, since the ownership of FTs can be related to the ownership of NFTs.

6 CONCLUSION AND FUTURE WORK

In this work, we analyzed the obtained data on both Fungible and non-Fungible tokens to understand the main distinguishing features and the possible correlations between economic features regarding social features of the Ethereum blockchain. We selected top ten PFP NFTs by highest trading volume. Different types of data including FT and NFT owners, their transactions and the owners' balances were explored. We leveraged "Moralis" API to get cross-chain data related to NFTs metadata, their ownership and transfer data. The data was then stored in a relational MySQL database for further analysis.

With regard to the three research questions, we have obtained extensive results that provide a deep insight into the NFT space. To this end, in the first research question, we found that participants in POAP events have had previous exposure to NFTs and are thus already familiar with the topic. Furthermore, we found that these participants perform NFT transfers both before and after attending their first event. In the second research question, we found that a large component exists in the transaction network that includes the most significant communities. We found this to be true for both FT and NFT transfers. Therefore, we can conclude that the most significant commuities are related and perform both FT and NFT transfers. In the third research question, we examined the relationship between economic features and social features and concluded that economic feautures in the form of FTs allow us to draw conclusions about the NFTs held. We confirmed this in the context of a user-based collaborative filtering recommendation system.

In addition, we addressed the classification of NFTs as PFPs at the beginning of the project. However, we had major problems with this, which is why we focused on a different research question and see this classification challenge as future work. Nevertheless, we would like to describe what we have already done here and what problems we faced. With the aim of designing a classifier that is able to distinguish traits from Metadata for predicting whether a NFT collection is categorized as PFPs (those collections with character-like features traits like eyes, nose, ears, hair, etc.), a list of 200 PFP collections was created. The metadata information was retrieved from these collections into our MySQL database in JSON format. The goal was to apply different machine learning algorithms to learn the aforementioned character-like features and deploy a classifier to figure out which features are predictive for PFP collections. To do this, we created a sample list of 1,000 NFTs derived from 200 different NFT collections (5 random NFTs were selected from each NFT collection). The next step was then to create a ground truth labeling method to label these samples manually by three different individuals and then, to apply majority voting system to judge whether these NFTs are categorized as PFPs. In contrast, a sample list of 1,000 non-PFPs was also created to create a balanced training dataset including both PFPs and non-PFPs. As next step, we tried to setup a machine learning task that uses these ground truth labels for training and to make a 10-fold cross validation to predict PFPs from NFT features. However, while we came to the conclusion that there are features that distinguish PFP Collections, but due to inconsistent and incomplete data we couldn't perform a machine learning Classification Algorithm. The main reasons for this problem were missing metadata information for many NFT Collections, missing key values in trait features and inconsistent traits identifiers like name, image and attribute.

As further work for future projects we see the possibility to use machine learning algorithms to develop another NFT recommender, which will give even better results. For example, multi-class classification could be used for this purpose. Our recommender could be used as the foundation for further research in this area.

REFERENCES

[1] Sarah Allen, Ari Juels, Mukti Khaire, Tyler Kell, and Siddhant Shrivastava. 2022. NFTs for Art and Collectables: Primer and Outlook. (2022).

[2] Simone Casale-Brunet, Paolo Ribeca, P Doyle, and Marco Mattavelli. 2021. Networks of Ethereum non-fungible Tokens: a graph-based analysis of the ERC-721 ecosystem. In *2021 IEEE International Conference on Blockchain (Blockchain)*. IEEE, 188–195.

[3] Usman W Chohan. 2021. Non-fungible tokens: Blockchains, scarcity, and value. *Critical Blockchain Research Initiative (CBRI) Working Papers* (2021).

[4] Michael Dowling. 2022. Is non-fungible token pricing driven by cryptocurrencies? *Finance Research Letters* 44 (2022), 102097.

[5] MohammadAmin Fazli, Ali Owfi, and Mohammad Reza Taesiri. 2021. Under the Skin of Foundation NFT Auctions. *arXiv preprint arXiv:2109.12321* (2021).

[6] Santo Fortunato. 2010. Community detection in graphs. *Physics reports* 486, 3-5 (2010), 75–174.

[7] Heng Luo, Changyong Niu, Ruimin Shen, and Carsten Ullrich. 2008. A collaborative filtering framework based on both local user similarity and global user similarity. *Machine Learning* 72, 3 (2008), 231–245.

[8] Matthieu Nadini, Laura Alessandretti, Flavio Di Giacinto, Mauro Martino, Luca Maria Aiello, and Andrea Baronchelli. 2021. Mapping the NFT revolution: market trends, trade networks, and visual features. *Scientific reports* 11, 1 (2021), 1–11.

[9] Mark EJ Newman and Michelle Girvan. 2004. Finding and evaluating community structure in networks. *Physical review E* 69, 2 (2004), 026113.

[10] Qin Wang, Rujia Li, Qi Wang, and Shiping Chen. 2021. Non-fungible token (NFT): Overview, evaluation, opportunities and challenges. *arXiv preprint arXiv:2105.07447* (2021).

[11] Zhi-Dan Zhao and Ming-Sheng Shang. 2010. User-based collaborative-filtering recommendation algorithms on hadoop. In *2010 third international conference on knowledge discovery and data mining*. IEEE, 478–481.

7 WORK DISTRIBUTION

Task	**Galm**	**Ostermaier**	**Agarwal**	**Garza**	**Omogha**	**Shekari**
Literature Review	X	X	X	X	X	X
Figma tool for Project management and documentation						X
Network Visualization with Gephi tool		X	X	X	X	X
Data Collection	X	X	X	X	X	X
Data Exploration	X	X	X	X	X	X
Classification of PFP	X	X	X	X	X	X
Mid-Term Presentation	X	X	X	X	X	X
End Term Presentation	X	X	X	X	X	X
Research Question #1	X	X	X	X		
Research Question #2		X	X	X	X	X
Research Question #3	X					
Writing Final Report: Introduction			X	X		X
Writing Final Report: Data Collection	X	X	X	X		
Writing Final Report: Research Question #1	X		X	X		
Writing Final Report: Research Question #2		X			X	
Writing Final Report: Research Question #3	X					X
Writing Final Report: Conclusion	X	X				X
Data Analysis: Ukraine Analysis			X	X		
Data Analysis: Research Question #1	X	X	X	X		
Data Analysis: Research Question #2	X	X	X	X		X
Data Analysis: Research Question #3	X					
Coding: Data Collection	X	X				
Coding: Research Question #1	X					
Coding: Research Question #2		X				
Coding: Research Question #3	X					